Proceedings of the Third International Conference on Trends in Electronics and Informatics (ICOEI 2019)
IEEE Xplore Part Number: CFP19J32-ART; ISBN: 978-1-5386-9439-8

Gold Price Prediction using Ensemble based Machine Learning Techniques

Manjula K A
Department of Computer Science
University of Calicut
manjulaka@gmail.com

Karthikeyan P
Department of Management Studies
Kannur University
karthik.dmsku@gmail.com

Abstract - This article is based on a study conducted to understand the relationship between gold price and selected factors influencing it, namely stock market, crude oil price, rupee dollar exchange rate, inflation and interest rate. Monthly price data for the period January 2000 to December 2018 was used for the study. The data was further split into two periods, period I from January 2000 to October 2011 during which the gold price exhibits a raising trend and period II from November 2011 to December 2018 where the gold price is showing a horizontal trend. Three machine learning algorithms, linear regression, random forest regression and gradient boosting regression were used in analyzing these data. It is found that the correlation between the variables is strong during the period I and weak during period II. While these models show good fit with data during period I, the fitness is not good during the period II. While random forest regression is found to have better prediction accuracy for the entire period, gradient boosting regression is found to give better accuracy for the two periods taken separately.

Keywords - Machine Learning, Regression, Prediction

I. INTRODUCTION

Savings and Investments form an integral part of everyone's life. Investments refer to the employment of present funds with an objective of earning a favourable return on it in future. In an economic sense, an investment can be considered as the purchase of assets that are not consumed today but are used in the future to create wealth. In finance, an investment is purchase of a monetary asset with the idea that the asset will provide income in the future or will later be sold at a higher price for a profit. The Indian economy being one of the fastest growing in the world has resulted in higher disposable income level and a plethora of investment avenues. There are a number of investment avenues available for investors, which includes stocks, deposits, commodities and real estate. Each of them differs in terms of risk and return characteristics. Gold is another asset which is being considered as an attractive investment avenue by many investors due to its increasing value and the area of usage. Investors preference for gold as a protective asset increases due to their negative expectations concerning the situation in the developed foreign exchange markets and the capital markets[1]. Gold is also considered to be "the asset of final instance" i.e. is the asset investors rely on, when the developed world capital markets are not capable to provide desirable profitability[2]. Thus it can be said that investors see gold as a tool to hedge against the fluctuations in other markets. Gold is a precious metal, so like any other goods, gold's price should depend on supply and demand. But, since gold is storable and the supply is accumulated over centuries, this year's production has little influence on its prices. Gold is used both as a commodity and as a financial asset. Gold behaves less like a commodity than long-lived assets such as stocks or bonds. Price of gold depends on a myriad of interrelated variables, including inflation rates, currency fluctuation and political turmoil[3].

This raising value of gold coupled with the volatilities and fall in prices of other markets like capital markets and real estate markets has attracted more and more investors towards gold as an attractive investment. But, of late price of gold is also witnessing high volatility and investments in gold are turning to be riskier. There is a fear as to whether these high prices are sustainable and when the prices would reverse. Eventhough there are a number of studies analyzing the correlation between the price of gold and some economic variables. It is still considered that a study to reveal the influence and impact of various macro-economic factors on the price of gold in the present situation will be helpful in determining the dynamic effects of these relationships. Thus this paper is aimed at studying the relationship between gold price and selected economic and market variable. Understanding such relationship will be helpful not only to monetary policymakers but also to investors, fund managers and portfolio managers to take better investment decisions in the market. Further this study uses three machine learning algorithms, linear regression, random forest regression and gradient boosting regression in analyzing these data. Comparison of these three methods will help us in identifying the accuracy of these methods under various conditions. This paper is structured with literature review in the next section followed by sections on data and methodology, results and discussion and conclusion.

II. LITERATURE REVIEW

There are many studies dealing with the price of gold in the literature. Although various different variables are used in these studies, it is observed that gold prices are regressed against USA dollar and stock return in general[4]. The relationship between other macroeconomic variable and gold prices has also been studied by many researchers. The relationship between gold price and prices of other commodities especially crude oil has also been extensively

studied. But the results from these studies are found to be contradicting. Some of the studies on the factors influencing gold price and various techniques used for studying these relationships are discussed in the following sections.

Lawrence[5] has found that there is no significant correlations between returns on gold and changes in certain macroeconomic variables such as inflation and GDP. He has also found that that gold returns are less correlated with returns on equity and bond indices than returns of other commodities. But, Sjaastad and Scacciavillani[6] reported that gold is a store of value against inflation and Baker and Van-Tassel [7] also have found that the price of gold depends on the future inflation rate. With respect to the relationship between gold price and inflation, based on the review of literature Hanan Naser[8] is of the opinion that historical studies with regards to the effectiveness of gold as a hedge against inflation are contradicting. Ismail et al.[3]have forecasted gold prices based on multiple economic factors such as commodity research bureau future index, USD/Euro foreign exchange rate, inflation rate, money supply, New York Stock Exchange Index; Standard and Poor 500 index, Treasury bill and USD index. The study finds that Commodity Research Bureau future index, USD/Euro foreign exchange rate, Inflation rate and money supply have a significant impact on gold price. Khaemusunun [9] has examined the impact of currencies of selected countries, Oil Prices and Interest Rate on the gold price. Hammoudeh et al. [10] conclude that there is an interdependent exist between the volatility of gold price and the exchange rate. Ai, et al.[11] report empirical evidence that the exchange rate relates to the gold price both in the long-run and short-run. Ewing and Malik [12] find evidence of volatility transmission between gold and oil future prices. Ghosh et al. [13] have concluded that gold prices are related with US Inflation level, interest rates and dollar exchange rate. They have also reported a long run relationship between gold prices and US Consumer Price Index as a result of the cointegration analysis. From the review of related literature, it can be concluded that the relationship between gold price and various factors considered to influence it are contradicting.

In studying volatility in gold price and the relationship with the factors considered to influence it, researchers have used a variety of techniques. Hossein and Abdolreza[14]have predicted the gold price by using artificial neural networks (ANN) and ARIMA model.Khaemusunun, (2009) predicts the Thai gold price by using Multiple Regression and ARIMA model. Toraman[4] has reported that various studies have been conducted using multivariate regression models to test the sensitivity of gold prices among various variables. In this regard Ismail et al.[3] have used multiple linear regression (MLR) model for forecasting the gold prices and are of the opinion that MLR model appeared to be useful for predicting the gold price. From the review of literature, it can be seen that multiple linear regression is widely used technique for understanding relationship among such variables.

III. DATA AND METHODOLOGY

Based on the review of literature five major factors that is considered to have influence on the gold price were identified. The factors that are considered for this study are stock market, crude oil price, rupee dollar exchange rate, inflation and interest rate. Values of Nifty500 index is taken as a representation of the stock prices. Nifty500 index represents the top 500 companies based on full market capitalization from the companies listed in National Stock Exchange. Inflation is represented by Consumer Price Index with base year as 2001. The term deposit rate for deposits above 1 year is taken as a representation for interest rate. Spot gold price in rupees per ounce is used to represent gold price. Monthly data from January 2000 to December 2018 were collected for all these variables. The databases of Centre for Monitoring Indian Economy were used for getting these data. There were 228 observations for each of these variables.

Machine learning algorithms were used to train and model the collected data. From the data collected, eighty percentage of the data was used for training and remaining twenty percentages for testing the model. The machine learning algorithms used in this study are linear regression, random forest regression and gradient boosting regression.

The statistical process for estimating the relationship between different variables is called regression analysis. Regression analysis is used to understand how the value of the dependent variable changes when one of the independent variables changes, while other variables are fixed.

Linear regression models with more than one independent variable are called multiple linear models. A representation of multiple linear regressions is where, Y is dependent variable and X_1, X_2 … are independent variables are as seen below.

$$Y = a + b_1*X_1 + b_2*X_2 + ... + b_p*X_p$$

As such, linear regression was developed in the field of statistics and is studied as a model for understanding the relationship between input and output numerical variables, but has been borrowed by machine learning. It is both a statistical algorithm and a machine learning algorithm now.

Decision trees can be used for various machine learning applications. Decision tree constructs a tree that is used for classification and regression. But trees that are grown really deep to learn highly irregular patterns tend to over-fit the training sets. A slight noise in the data may cause the tree to grow in a completely different manner.

Ensemble methods are techniques that create multiple models and then combine them to produce improved results. Ensemble methods usually produces more accurate solutions than a single model would. The Random Forests and gradient-boosted trees are called ensemble methods. Ensemble methods take multiple weak learners, such as decision trees, and construct a strong learner from them such as random forest.

The Random Forests and gradient-boosted trees both can be used for classification and regression tasks.

Decision trees can be used for various machine learning applications. Decision tree constructs a tree that is used for classification and regression. But trees that are grown really deep to learn highly irregular patterns tend to over-fit the training sets. A slight noise in the data may cause the tree to grow in a completely different manner.

A Random Forest is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and a technique called Bootstrap Aggregation, commonly known as bagging. The basic idea behind this is to combine multiple decision trees in determining the final output rather than relying on individual decision trees. The Random Forest uses bootstrapping on Decision Trees to reduce the variance while maintaining the low bias that is resulted from a Decision Tree model. A Random Forest algorithm has the following advantages when compared to most of the other algorithms - The overfitting problem will never come when we use the random forest algorithm in any classification problem. The same random forest algorithm can be used for both classification and regression task. And, the random forest algorithm can be used for feature engineering for identifying the most important features out of the available features from the training dataset.

Gradient boosting is a machine learning technique which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function. Gradient boosting trees are better than random forests in many situations, but they are prone to overfitting in some situations. However, there are strategies to overcome the same and build more generalized trees using a combination of parameters like learning rate (shrinkage) and depth of tree. Generally, these two parameters are kept on the lower side to allow for slow learning and better generalization.

In this study these machine learning algorithms (Linear Regression, Random Forest Regression and Gradient Boosting Regression) are implemented using python. The prediction accuracy of the regression methods used were measured using Mean Squared Error (MSE), Root Mean Square Error (RMSE) and Mean Absolute Error (MAE)

IV. RESULTS AND DISCUSSION

The monthly gold price for the period 2000-2018 was plotted (Fig.1), it can be seen that gold price exhibits two characteristic trends during this period. From January 2000 to October 2011, the price is seen raising with an upward trend, but after that the price shows a horizontal trend (Table 1). Period from January 2000 to October 2011 is considered as period I and from November 2011 to December 2018 is considered as period II for this study.

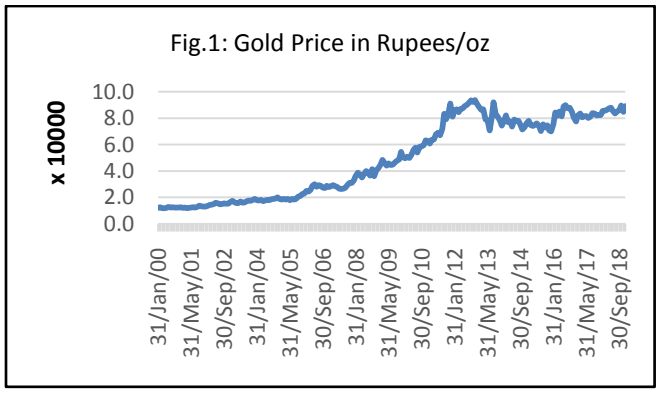


Table 1: Descriptive statistics

Period	Mean	Std. Dev.	Coeff. of Variation
Jan 2000 - Dec 2018	49627	29304	59.04
Jan 2000- Oct 2011 (period I)	26235	13643	52
Nov 2011 - Dec 2018 (period II)	82163	6047	7.36

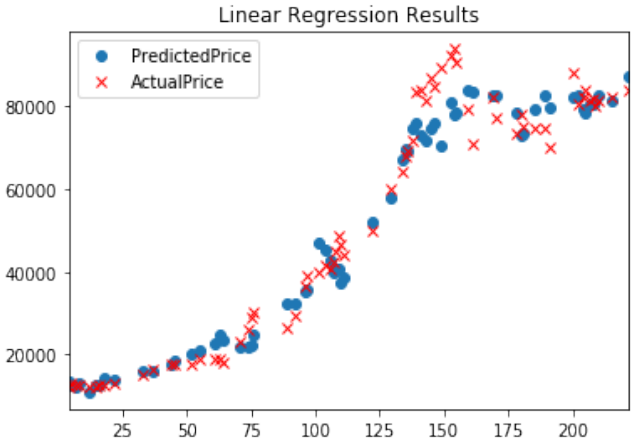
From Table 1 it can be inferred that the gold price for the period under study has been increasing, but the while during the period I there was a rapid increase in price, the price is showing a horizontal movement during period II. Similarly, there has been a drastic decrease in volatility measured through standard deviation and in coefficient of variation during period II compared to Period I.

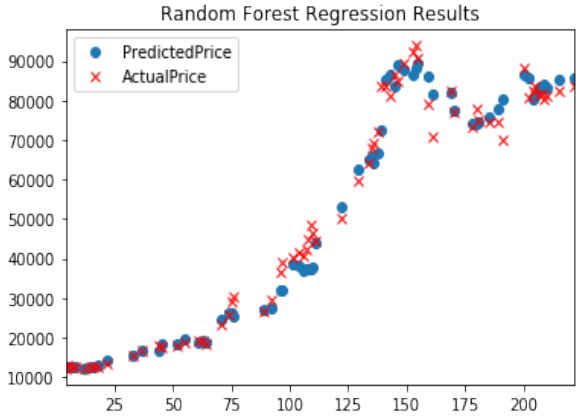
From the result of correlation done on this data (Table 2), it is seen that for the entire period of study, there is strong positive correlation between gold price and other variables except interest rate. During the period I, while the gold price exhibits a strong correlation with stock market, crude oil and inflation, there is very weak correlation with exchange rate. This may be due to the fact that Rupee- US Dollar exchange rate has mostly stable during this period. But in the case of period II, it is seen that gold prices exhibit very weak negative correlation with all these factors. This may be due to the fact that the gold price during this period is not increasing compared to the period I.

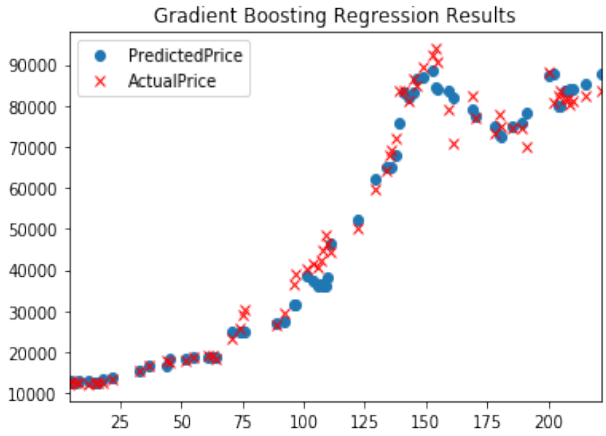
Table 2: Correlation between Gold and other factors			
	Jan 2000- Dec 2018	Jan 2000 - Oct 2011	Nov 2011 - Dec 2018
NIFTY 500	0.855107	0.840182	-0.060482
Crude Oil	0.60708	0.806525	0.251313
Exch rate	0.770875	-0.00943	-0.183818
Inflation	0.939483	0.986389	-0.129957
Interest rate	0.128474	0.198394	-0.098324

Table 3: Prediction Accuracy				
		Jan 2000 - Dec 2018	Jan 2000 - Oct 2011	Nov 2011 - Dec 2018
Linear Regression	R^2	0.957	0.9617	0.1549
	MAE	4384.43	2075.92	4138.1
	MSE	35465098.85	10627560.74	27918843.05
	RMSE	5955.25	3259.99	5283.82
Random Forest Regression	R^2	0.9802	0.9624	0.5586
	MAE	2808.44	1891.24	2857.05
	MSE	16294903.01	10433534.92	14580317.81
	RMSE	4036.69	3230.09	3818.41
Gradient Boosting Regression	R^2	0.9786	0.979	0.6288
	MAE	3007.8	1529.11	2728.56
	MSE	17609001.94	5831754.47	12264286.33
	RMSE	4196.3	2414.9	3502.04

Fig. 2: Test Results for the period 2000-2018 for the three Algorithms







Comparing the prediction accuracy of the three models (Table 3), it is seen that the R squared value in the case of the entire period and period I is very high (more than 95%). This denotes that models from all the three algorithms fit the data very well. While in the case of the entire period, random forest regression is found to have the highest fit, in the case of period I gradient boosting regression is found to have the highest fit (Fig.2). The R squared value in case of period II is found to be low, among the methods the value for regression is very low and gradient boosting regression has the highest fit for this period. This may be due to the fact that correlation between gold price and other factors are high for the entire period and period I, but for period II the correlation is low.

With respect to the prediction accuracy using Mean Squared Error (MSE), Root Mean Square Error (RMSE) and Mean Absolute Error (MAE), random forest has the lowest values for the entire period and gradient boosting regression has the lowest values for the other two periods. While for a longer period random forest regression is found to be performing better, in the case of shorter period gradient boosting regression is better.

Based on the results from these analyses, it may be inferred that there has been a change in the trend in the gold price

during the period considered for this study. In such situations where there is change in the trend of the dependent variable and no significant changes in the trend of independent variables, the accuracy of various methods may differ. Hence the model used should depend on the relationship between the variables used in the study.

V. CONCLUSION

This study was conducted to understand the relationship between gold price and selected factors influencing its price, namely stock market, crude oil price, rupee dollar exchange rate, inflation and interest rate. Monthly price data for the period January 2000 to December 2018 was used for the study. The data was further split into two periods, period I from January 2000 to October 2011 during which period the gold price exhibits a raising trend and period II from November 2011 to December 2018 where the gold price is showing a horizontal trend. Three machine learning algorithms, linear regression, random forest regression and gradient boosting regression were used in analyzing these data. It is found that the correlation between the variables is strong during the period I and weak during period II. While these models show good fit with data during period I, the fitness is not good during the period II. Random forest regression is found to have better prediction accuracy for the entire period and gradient boosting regression is found to give better accuracy for the two period taken separately. It is concluded that machine learning algorithms are very useful in such analysis, but the characteristics of the data influences their accuracy. Further research with such data and different techniques may be conducted for better understanding of the performance of these techniques.

REFERENCES

[1] W. Du and J. Schreger, "Local Currency Sovereign Risk," Social Science Research Network, Rochester, NY, SSRN Scholarly Paper ID 2976788, Dec. 2013.

[2] J. Jagerson and S. W. Hansen, "All about investing in gold", McGraw-Hill Publishing, 2011.

[3] Z. Ismail, A. Yahya, and A. Shabri, "Forecasting gold prices using multiple linear regression method," Am. J. Appl. Sci., vol. 6, no. 8, p. 1509, 2009.

[4] C. Toraman, Ç. Basarir, and M. F. Bayramoglu, "Determination of factors affecting the price of gold: A study of MGARCH model," Bus. Econ. Res. J., vol. 2, no.4, p. 37, 2011.

[5] C. Lawrence, "Why is gold different from other assets? An empirical investigation," Lond. UK World Gold Council., 2003.

[6] L. A. Sjaastad and F. Scacciavillani, "The price of gold and the exchange rate," J. Int. Money Finance, vol. 15, no.6, pp. 879–897, 1996.

[7] S. A. Baker and R. C. Van Tassel, "Forecasting the price of gold: A fundamentalist approach," Atl. Econ. J., vol. 13, no. 4, pp. 43–51, 1985.

[8] H. Naser, "Can Gold Investments Provide a Good Hedge Against Inflation? An Empirical Analysis," Int. J. Econ. Financ. Issues, vol. 7, no. 1, pp. 470–475, 2017.

[9] P. Khaemasunun, "Forecasting Thai gold prices," Available Http://www Wbiconpro Com3-Pravit. Pdf Acess, vol. 2, 2014.

[10] S. M. Hammoudeh, Y. Yuan, M. McAleer, and M. A. Thompson, "Precious metals–exchange rate volatility transmissions and hedging strategies," Int. Rev. Econ. Finance, vol. 19, no. 4, pp. 633–647, 2010.

[11] A. Han, K. K. Lai, S. Wang, and S. Xu, "An interval method for studying the relationship between the Australian dollar exchange rate and the gold price," J. Syst. Sci. Complex., vol. 25, no. 1, pp. 121–132, 2012.

[12] B. T. Ewing and F. Malik, "Volatility transmission between gold and oil futures under structural breaks," Int. Rev. Econ. Finance, vol. 25, pp. 113–121, 2013.

[13] D. Ghosh, E. J. Levin, P. Macmillan, and R. E. Wright, Gold as an inflation hedge?," Stud. Econ. Finance, vol. 22, no. 1, pp. 1–25, 2004.

[14] H. Mombeini and A. Yazdani-Chamzini, "Modeling gold price via artificial neural network," J. Econ. Bus. Manag., vol. 3, no. 7, pp. 699–703, 2015.